Statistical Analysis of Economic Variables and the Mortgage Rate in the United States: Exploring Influential Factors and Predictive Models

ECON 400: Econometrics- Bellevue College

Project Paper

By Candice Wu

I. INTRODUCTION

The purpose of this statistical analysis is to examine the relationship between various economic variables and the 30-Year Fixed Rate Mortgage Average in the United States (MORTGAGE30US). In this analysis, I aim to investigate the impact of these independent variables on the mortgage rate and determine which variables have a significant relationship with MORTGAGE30US. Additionally, I will address challenges such as multicollinearity and skewness in the dataset, performing appropriate techniques to mitigate their effects. The findings of this analysis will contribute to a better understanding of the factors influencing mortgage rates in the United States.

II. II DATA

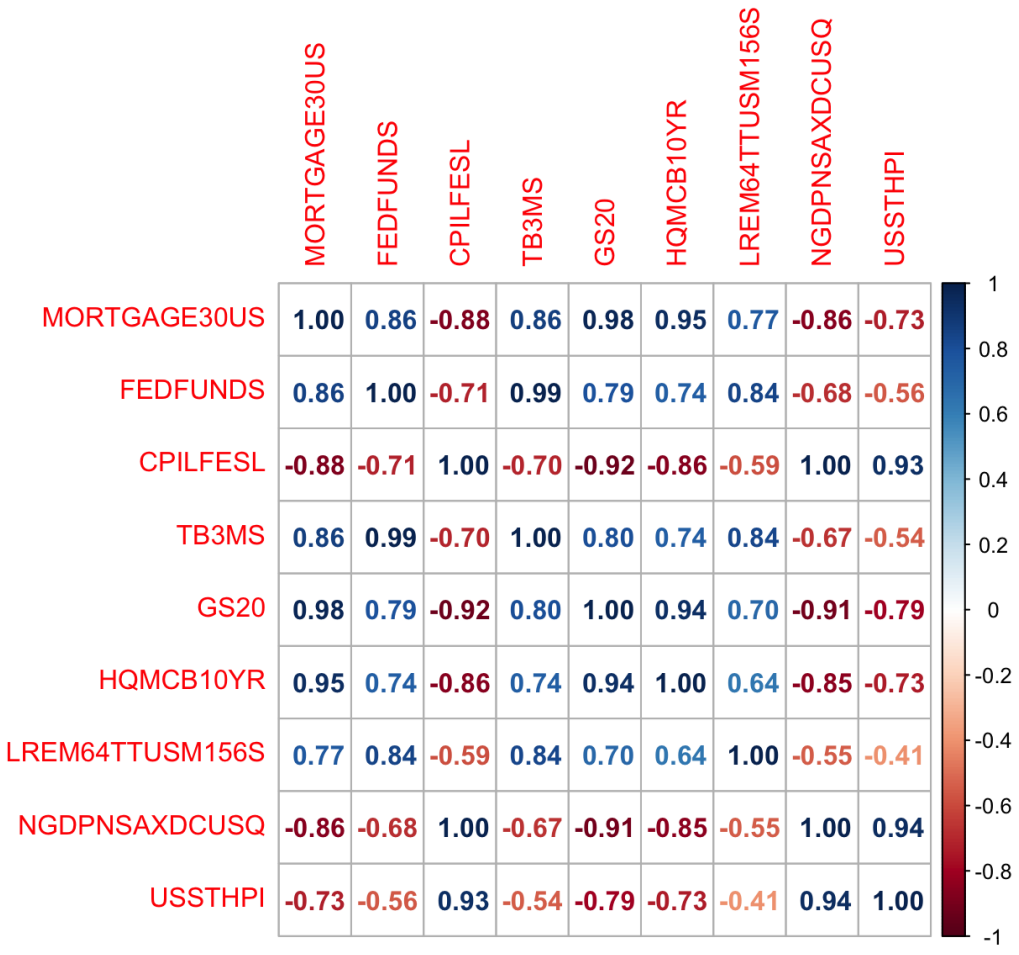
A. About The Dataset

The dataset used in this analysis is extracted from the Federal Reserve's Economic Data (FRED) database, covering the period from October 1, 1993, to October 31, 2022. The dataset comprises 349 observations and includes 9 variables. The dependent variable is MORTGAGE30US, and the independent variables are FEDFUNDS, CPILFESL, NGDPNSAXDCUSQ, TB3MS, GS20, USSTHPI, HQMCB10YR, and LREM64TTUSM156S.

B. Variables

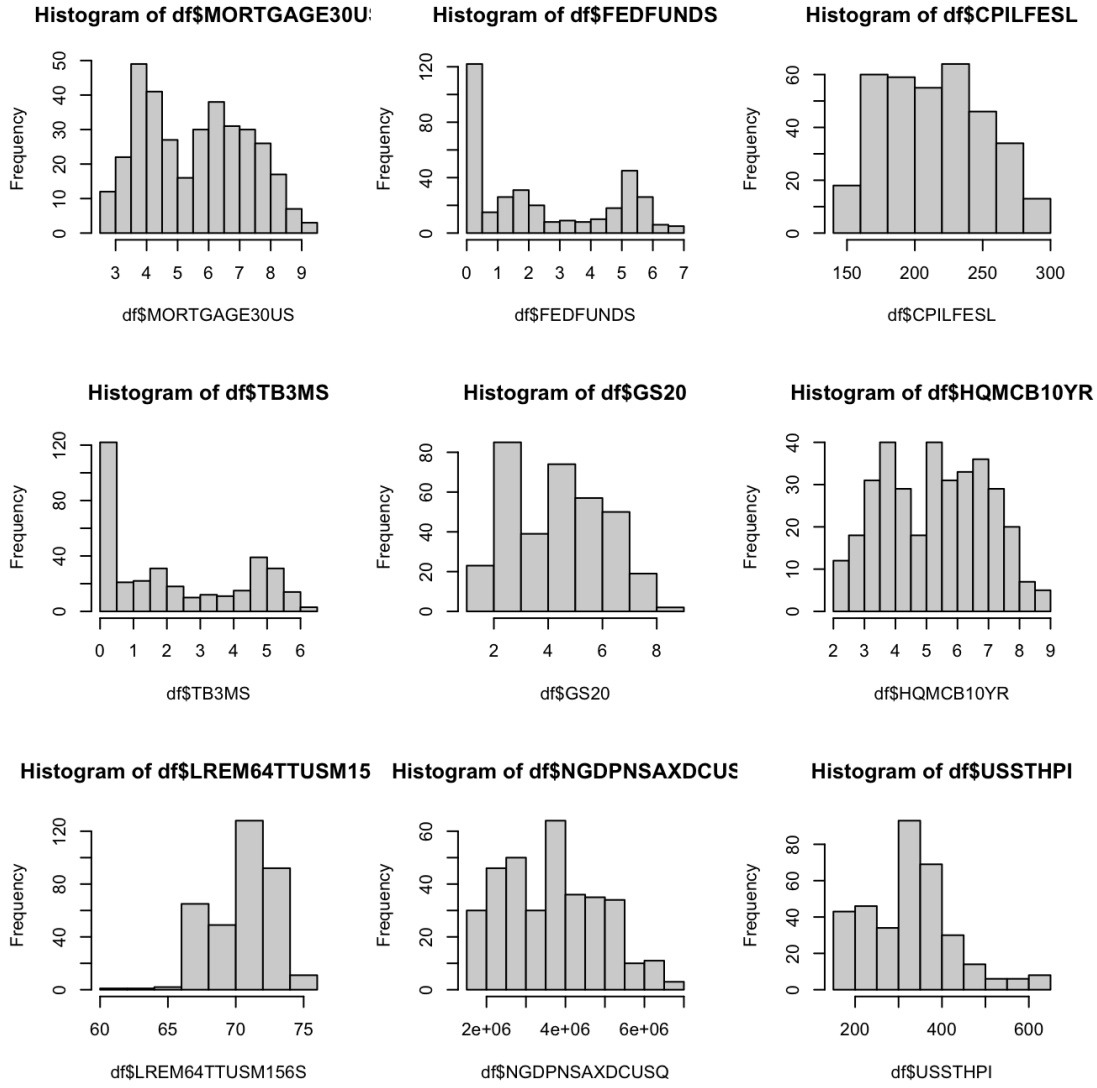
MORTGAGE30US: 30-Year Fixed Rate Mortgage Average in the United States
FEDFUNDS: Federal Funds Effective Rate
CPILFESL: Consumer Price Index for All Urban Consumers: All Items Less Food and Energy in U.S. City Average
NGDPNSAXDCUSQ: Nominal Gross Domestic Product for United States
TB3MS: 3-Month Treasury Bill Secondary Market Rate, Discount Basis
GS20: Market Yield on U.S. Treasury Securities at 20-Year Constant Maturity, Quoted on an Investment Basis
USSTHPI: All-Transactions House Price Index for the United States
HQMCB10YR: 10-Year High Quality Market (HQM) Corporate Bond Spot Rate
LREM64TTUSM156S: Employment Rate: Aged 15-64: All Persons for the United States

C. Data Visualization



When analyzing a dataset for statistical purposes, it is vital to examine the correlation matrix of variables as it provides valuable insights into the strength, direction, and significance of relationships between variables. The correlation coefficient ranges from -1 to 1, where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values close to 0 indicate little to no correlation.

Based on the plot of the correlation matrix provided above, there are many pairs of variables that are highly correlated with each other. It also indicates the presence of multicollinearity in the dataset.



Furthermore, it is essential to assess the distribution of each variable to ensure the fulfillment of assumptions required by many statistical functions, particularly the assumption of normality or near-normality. Additionally, skewness is a measure of the asymmetry of a random variable's distribution around its mean and can help identify deviations from a symmetrical distribution. If the histogram has a long tail on the right side and the majority of the data is concentrated on the left side, it suggests a right-skewed distribution. Moreover, if the histogram has a long tail on the left side and

the majority of the data is concentrated on the right side, it suggests a left-skewed distribution.
In this dataset, variables such as FEDFUNDS, TB3MS, and USSTHPI demonstrate right-skewed distributions, while LREM64TTUSM156S displays a left-skewed distribution

III. METHODS and MODEL

A. Applying Transformation

```
> skewness(df)
    MORTGAGE30US        FEDFUNDS         CPILFESL
     0.154982613     0.444198338     0.212395757
            TB3MS            GS20        HQMCB10YR
     0.419375821     0.108274727    -0.005290681
 LREM64TTUSM156S    NGDPNSAXDCUSQ          USSTHPI
    -0.634892449     0.323665697     0.675757279
```

Given that both USSTHPI and LREM64TTUSM156S exhibit absolute skewness values between 0.5 and 1, indicating moderate skewness, it is appropriate to apply the square root transformation method. This method can help achieve a more symmetrical distribution and meet the assumptions.

B. Handling Multicollinearity Issue

To address the presence of multicollinearity in the dataset, one approach is to check the Variance Inflation Factor (VIF) and drop variables with high VIF values. “As an arbitrary rule of thumb, it is often suggested that the VIF should not exceed 10.” (Dr. Lawrence, wk3)

By dropping variables with high VIF values one at a time, the final set of independent variables in the model are as follows:

FEDFUNDS with VIF of 4.558909
USSTHPI with VIF of 2.255771
HQMCB10YR with VIF of 3.373104
LREM64TTUSM156S with VIF of 3.548210

C. Model

Three models will be conducted for analysis. The first model is an Ordinary Least Squares (OLS) model with the dependent variable (MORTGAGE30US) and independent variables including FEDFUNDS, CPILFESL, NGDPNSAXDCUSQ, TB3MS, GS20, USSTHPI, HQMCB10YR, and LREM64TTUSM156S. The second model is also an OLS model with the dependent variable and independent variables consisting of FEDFUNDS, USSTHPI, sqrt_HQMCB10YR, and sqrt_LREM64TTUSM156S. In this model, the independent variables sqrt_HQMCB10YR and sqrt_LREM64TTUSM156S are the square root transformations of HQMCB10YR and LREM64TTUSM156S. The last model will utilize Weighted Least Squares (WLS) regression. It will have the dependent variable and independent variables, including FEDFUNDS, USSTHPI, sqrt_HQMCB10YR, and sqrt_LREM64TTUSM156S.

IV. RESULTS and INTERPRETATION

A. Full model

```
Call:
lm(formula = MORTGAGE30US ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81245 -0.10877 -0.00577  0.10142  0.82405

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.343e+00  8.658e-01  -3.861 0.000135 ***
FEDFUNDS        -1.319e-01  4.951e-02  -2.665 0.008071 **
CPILFESL         8.014e-03  4.872e-03   1.645 0.100880
TB3MS            2.817e-01  5.356e-02   5.260 2.56e-07 ***
GS20             5.241e-01  3.075e-02  17.042  < 2e-16 ***
HQMCB10YR        3.099e-01  1.946e-02  15.930  < 2e-16 ***
LREM64TTUSM156S  5.006e-02  8.202e-03   6.103 2.83e-09 ***
NGDPNSAXDCUSQ   -2.227e-07  1.403e-07  -1.587 0.113365
USSTHPI          7.257e-04  3.552e-04   2.043 0.041824 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1891 on 340 degrees of freedom
Multiple R-squared:  0.9872,    Adjusted R-squared:  0.9869
F-statistic:  3281 on 8 and 340 DF,  p-value: < 2.2e-16
```

The coefficients for FEDFUNDS, TB3MS, GS20, USSTHPI, HQMCB10YR, and LREM64TTUSM156S are statistically significant ($p < 0.05$), indicating a significant relationship between these variables and the MORTGAGE30US. This suggests that changes in these variables are associated with changes in the mortgage rate.

The F-statistic is 3281 with a p-value of < 2.2e-16, indicating that the overall model is highly statistically significant. This means that the independent variables, as a group, have a significant impact on predicting the mortgage rate.

However, in situations where variables demonstrate high correlation, it can be challenging to determine the individual effects of each variable on the dependent variable. Moreover, the presence of skewness in the data has the potential to violate the assumptions of linear regression and affects the accuracy of the results.

B. Best Model

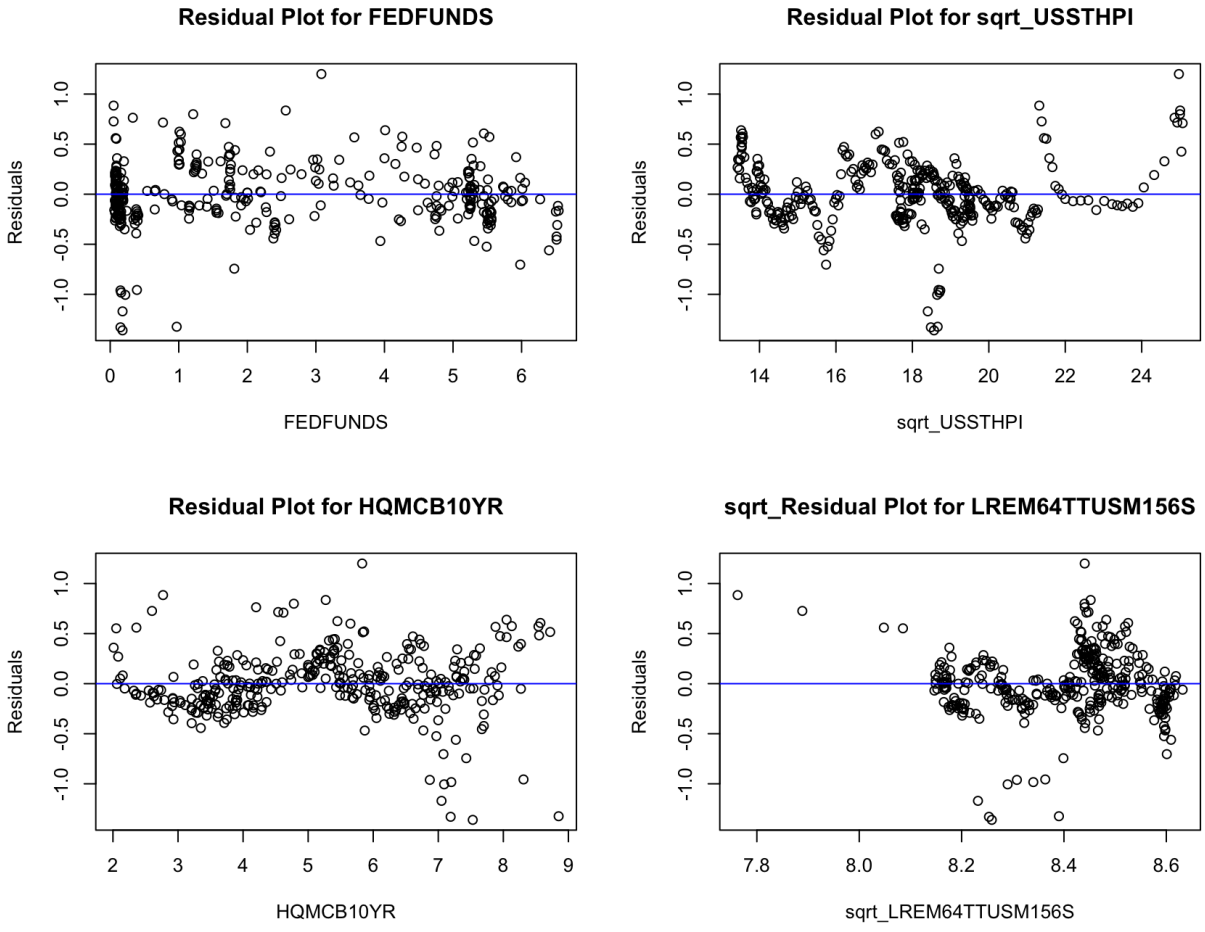
```
Call:
lm(formula = MORTGAGE30US ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3601 -0.1726 -0.0156  0.1754  1.2003

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -10.137416   1.796394  -5.643 3.49e-08 ***
FEDFUNDS              0.167385   0.016575  10.099  < 2e-16 ***
HQMCB10YR             0.629939   0.019293  32.651  < 2e-16 ***
sqrt_USSTHPI         -0.055853   0.009865  -5.662 3.17e-08 ***
sqrt_LREM64TTUSM156S  1.545395   0.220606   7.005 1.30e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3214 on 344 degrees of freedom
Multiple R-squared:  0.9626,    Adjusted R-squared:  0.9622
F-statistic:  2214 on 4 and 344 DF,  p-value: < 2.2e-16
```

The coefficients for FEDFUNDS, HQMCB10YR, sqrt_USSTHPI, and sqrt_LREM64TTUSM156S are statistically significant ($p < 0.05$), indicating a significant

relationship between these variables and the MORTGAGE30US. This suggests that these variables have a meaningful impact on the mortgage rate.

The F-statistic is 2214 with a p-value of < 2.2e-16, indicating that the overall model is highly statistically significant. This means that the independent variables, have a significant impact on predicting the mortgage rate.
The R-squared value of 0.9626 indicates that approximately 96.26% of the variance in the mortgage rate can be explained by the independent variables in the model. This implies that the model provides a strong fit to the data.

1. the residuals against each of the x-variables



Based on the residual plots, the points appear to be randomly scattered around the residual = 0 line. This indicates that the assumptions of linearity and constant variance are reasonably met, suggesting that a linear model is appropriate for modeling this data.

2. White's Test

```
> bptest(model2, ~ fitted(model2) + I(fitted(model2)^2) )

        studentized Breusch-Pagan test

data:  model2
BP = 3.7558, df = 2, p-value = 0.1529
```

Based on the p-value being less than 0.05, we reject the null hypothesis. Therefore, we have sufficient evidence to conclude that heteroscedasticity is present in the regression model. This implies that the assumption of equal variance of the error terms is violated.

In the presence of heteroscedasticity, the OLS estimator may still be unbiased, but the usual F-test and t-tests may not be valid. Therefore, caution should be exercised when interpreting the significance of individual coefficients. It may be necessary to employ alternative statistical techniques or robust standard errors to account for the heteroscedasticity and obtain valid inferences from the model.

3. Robust Standard Errors

```
t test of coefficients:

                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)         -10.137416   2.528454 -4.0093 7.471e-05 ***
FEDFUNDS              0.167385   0.025136  6.6593 1.091e-10 ***
HQMCB10YR             0.629939   0.042837 14.7053 < 2.2e-16 ***
sqrt_USSTHPI         -0.055853   0.017664 -3.1620  0.001706 **
sqrt_LREM64TTUSM156S  1.545395   0.309828  4.9879 9.704e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By using robust standard errors, which are designed to handle heteroscedasticity, all the coefficients in the model are statistically significant in explaining the variation in the dependent variable. This indicates that these variables have a meaningful impact on the outcome variable, even when accounting for heteroscedasticity.

C. WLS Model

```
Call:
lm(formula = MORTGAGE30US ~ ., data = df, weights = wt)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.8911 -0.7275 -0.0972  0.7731  4.8471

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -8.361199   1.625883  -5.143 4.56e-07 ***
FEDFUNDS              0.180713   0.016320  11.073  < 2e-16 ***
HQMCB10YR             0.645064   0.019041  33.877  < 2e-16 ***
sqrt_USSTHPI         -0.044049   0.009677  -4.552 7.38e-06 ***
sqrt_LREM64TTUSM156S  1.295562   0.200950   6.447 3.86e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.403 on 344 degrees of freedom
Multiple R-squared:  0.9629,    Adjusted R-squared:  0.9625
F-statistic:  2235 on 4 and 344 DF,  p-value: < 2.2e-16
```

All variables in the model have p-values less than 0.05, indicating their statistical significance in explaining the variation in the mortgage rate. The F-statistic of 2235 and a p-value of < 2.2e-16 indicate the overall high statistical significance of the model.

The R-squared value of 0.9629 indicates that approximately 96.29% of the variance in the MORTGAGE30US can be explained by the independent variables. The adjusted R-squared value of 0.9625 takes into account the number of predictors in the model, providing a more conservative estimate of the proportion of variance explained.

V. CONCLUSION

In conclusion, this statistical analysis examined the relationship between various economic variables and the 30-Year Fixed Rate Mortgage Average in the United States. Through the evaluation of three different models, I identified key variables that have a significant impact on the mortgage interest rate.

The results of the best model revealed that the FEDFUNDS, HQMCB10YR, sqrt_USSTHPI, and sqrt_LREM64TTUSM156S were statistically significant predictors of MORTGAGE30US. These findings suggest that changes in these variables are associated with changes in the mortgage interest rate.

Moreover, the models demonstrated a high overall statistical significance, indicating that the independent variables collectively have a strong impact on predicting the mortgage interest rate. The high R-squared values of the models indicate a good fit to the data, with approximately 96% of the variance in the MORTGAGE30US explained by the independent variables.

These findings provide valuable insights into the factors influencing the mortgage rate and can assist in making informed decisions in the real estate and financial sectors. However, it is important to acknowledge the limitations of the models, such as the presence of heteroscedasticity and the need for further robustness checks. Further research and analysis are recommended to validate and enhance the findings presented in this study.

References:

"Economic Theory Blog." Robust Standard Errors. Economic Theory Blog, 7 Aug. 2016, https://economictheoryblog.com/2016/08/07/robust-standard-errors/#:~:text=%E2%80%9CRobust%E2%80%9D%20standard%20errors%20is%20a,linear%20unbiased%20estimator%20(BLUE).

"How to Interpret a Residual Plot - Explanation." Study.com, Study.com, n.d. Accessed 26 May 2023. https://study.com/skill/learn/how-to-interpret-a-residual-plot-explanation.html.

FRED (Federal Reserve Economic Data)." Federal Reserve Bank of St. Louis, fred.stlouisfed.org/

"Weighted Least Squares in R - Statology." Statology, 2023, www.statology.org/weighted-least-squares-in-r/

"Skewness and Kurtosis in R - Statology." Statology, 2023, www.statology.org/skewness-kurtosis-in-r/

Wk3-Collinearity

WK6-Hsk

Stargazer:

	Dependent variable:		
	MORTGAGE30US		
	(1)	(2)	(3)
FEDFUNDS	-0.134***	0.167***	0.181***
	(0.050)	(0.017)	(0.016)
CPILFESL	0.008*		
	(0.005)		
NGDPNSAXDCUSQ	-0.00000		
	(0.00000)		
TB3MS	0.286***		
	(0.054)		
GS20	0.531***		
	(0.031)		
HQMCB10YR	0.309***	0.630***	0.645***
	(0.020)	(0.019)	(0.019)
sqrt_USSTHPI	0.020	-0.056***	-0.044***
	(0.013)	(0.010)	(0.010)
sqrt_LREM64TTUSM156S	0.830***	1.545***	1.296***
	(0.136)	(0.221)	(0.201)
Constant	-7.030***	-10.137***	-8.361***
	(1.321)	(1.796)	(1.626)
Observations	349	349	349
R2	0.987	0.963	0.963
Adjusted R2	0.987	0.962	0.963
Residual Std. Error	0.190 (df = 340)	0.321 (df = 344)	1.403 (df = 344)
F Statistic	3,253.459*** (df = 8; 340)	2,213.558*** (df = 4; 344)	2,234.684*** (df = 4; 344)
Note:			*p<0.1; **p<0.05; ***p<0.01